ON-LINE GENRE CLASSIFICATION OF TV PROGRAMS USING AUDIO CONTENT

Samuel Kim[1]*, Panayiotis Georgiou*[2]*, Shrikanth Narayanan*[2]

[1] DSP Lab., Yonsei University, Seoul, Korea
[2] SAIL Lab., University of Southern California, Los Angeles, California, USA.
samuel.kim@dsp.yonsei.ac.kr

ABSTRACT

Automatic genre classification of TV programs can benefit users in various ways such as allowing for rapid selection of multimedia content. In this paper, we introduce an on-line method that can classify genres of TV programs using audio content. We deploy an acoustic topic model (ATM) which was originally designed to capture contextual information embedded within audio segments. With a dataset based on RAI content, we perform both on-line and off-line classification; we segment audio signals with a fixed length and feed into the system for on-line classification tasks, while we use whole audio signals for off-line tasks. The off-line experimental results suggest that the proposed method using audio content yields competitive performance with conventional methods using audio-visual features and outperforms conventional audio-based approaches. The on-line results show promising results in classifying genre of TV programs with short segments and also suggest that ATM performs better than conventional GMM method if the length of audio segments is longer (>1 second).

1. INTRODUCTION

Television is one of the most used electronic devices in our daily life. Beyond simple transmission of multimedia content from broadcast service providers to end users, recent advances in technologies related to content analysis enable TV to assume various functionalities, such as summarization and scene browsing [1, 2].

Along this line, there is a growing body of interest in classifying the genres of TV programs and videos to benefit users in various ways [3]. While videos available on the internet often have corresponding textual information (e.g., titles or tags), TV programs have, if any, limited textual information. Therefore, studies on TV programs mainly use signal-based features extracted from video and audio signals followed by supervised classifiers that can learn the patterns of the features. In particular, Montagnuolo *et al.* [4, 5] and Ekenel *et al.* [6] have shown promising results using various audio-visual features and machine learning techniques. In their experiments, they used the RAI dataset (an Italian TV broadcast database) that Montagnuolo *et al.* collected and distributed [4]; we will employ the same dataset in our work. Through the use of the same dataset, the above studies and our current study provide a single benchmark to the research community.

The contributions of this work are two-fold. First, we investigate audio content toward enabling genre classification. The rationale behind this is that audio content in a TV program includes abundant information about the genre of the corresponding TV program but there still remain open challenges in dealing with audio signals (potentially contributing to overall performance improvement when combined with visual features). The challenges are often related to ambiguities in a heterogeneous mixture. Generic audio signals, especially ones in TV programs, are generally heterogenous mixtures of several sound sources and this heterogeneity leads to the importance of exploiting context in the interpretation of sounds; perceptively similar acoustic content may lead to different semantic interpretation depending on the evidence provided by the co-occurring acoustic sources. The topic of computing context-dependencies in generic audio signals has been well studied. The closest work may be found in [7] where Lee *et al.* have used probabilistic latent semantic analysis (pLSA) [8] in consumer video classification tasks with a set of semantic concepts. Only with audio information from video clips, they decomposed the GMM histograms of feature vectors using pLSA to remove redundant structure and demonstrated promising performance in classifying video clips. We have further developed the idea and proposed acoustic topic models (ATM) using latent Dirichlet allocation (LDA) [9, 10, 11] and supervised LDA [12, 13, 14], and showed promising results in classifying generic audio signals with respect to semantic labels and onomatopoeic labels.

Secondly, we attempt on-line genre classification tasks which do not require any prior segmentation information. Note that classification tasks using already-segmented clips (off-line) implicitly assume that the system somehow knows when to make a decision, i.e., the starting and ending points, which is not trivial in practice. Furthermore, in some applications that require real-time processing of TV programs (e.g., real-time picture mode selection or real-time audio configuration adjustment), low-latency strategies are required. Therefore, here we examine the classification task in an on-line scenario, to detect genre information using fixed length segments, which are much shorter than the original clip (in particular, we attempt various lengths of segments from 0.5 seconds to 6 seconds).

In this work, we study how context modeling through ATM can enable genre classification for different audio segment length, either on-line or off-line. The tasks are evaluated with the same database, under similar conditions as [4] so that the results can be easily compared with other benchmarks.

2. ACOUSTIC TOPIC MODEL

Let V be the number of acoustic words (basic units of representation, which will be defined later in this section) in a dictionary and w be a V-dimensional vector whose elements are zero except for the corresponding acoustic word index. Assume that an audio clip consists of N words and can be represented as $\mathbf{d} = \{w_1, w_2, \cdots, w_i, \cdots, w_N\}$ where w_i is the i-th word in the document. Let the dataset consist of M documents and be represented as $S = \{\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_M}\}$. Suppose there are k latent acoustic topics and each word w_i is generated by its corresponding topic. The generative process can be illustrated as in Fig. 1 and described as

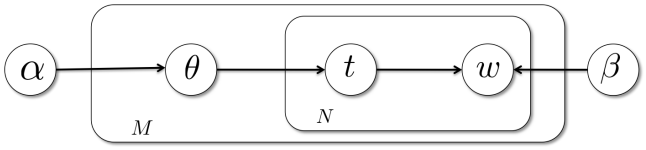


Fig. 1. Graphical representation of the topic model using LDA.

follows (see [10, 11] for more details);

1. For each document $\mathbf{d}$ in data set S
   (a) Choose the topic distribution $\theta \sim Dir(\alpha)$, where $Dir(\cdot)$ and α represent a Dirichlet distribution and its Dirichlet coefficient, respectively.
2. For each word w_i in document $\mathbf{d}$,
   (a) Choose a topic $t_i \sim$ Multi(θ), where t_i is the topic that corresponds with the word w_i and $Multi(\cdot)$ represents a multinomial distribution.
   (b) Choose a word w_i with a probability $p(w_i|t_i, \beta)$, where β denotes a $k \times V$ matrix whose elements represent the probability of a word with a given topic, i.e., $\beta_{nm} = p(w_i = m|t_i = n)$.

To define acoustic words, one may come up with various methodologies to transform an audio signal to a sequence of word-like units that represent specific characteristics of the audio signal. In this paper, for simplicity, we use conventional vector quantization (VQ) to derive the acoustic words from the feature vectors (mel frequency cepstral coefficients, MFCC) which provide spectral parametrization of the audio signal considering human auditory properties and have been widely used in many sound related applications, such as speech recognition and audio classification [15]. We use 20 ms hamming windows with 50% overlap to extract 12-dimensional feature vectors. With a given set of MFCCs, we derive a dictionary of acoustic words using the *Linde-Buzo-Gray Vector Quantization* (LBG-VQ) algorithm [16]. The rationale is to cluster audio segments which have similar acoustic characteristics and to represent them as discrete indexed numbers. Once the dictionary is built, the extracted acoustic feature vectors from sound clips can be mapped to acoustic words by choosing the closest word in the dictionary. We set the number of words in the dictionary, i.e., vocabulary size V, empirically (in this work, 2048).

To deploy the proposed acoustic topic model (ATM) in the framework of TV genre classification, we adopt a two-step learning strategy: an unsupervised acoustic topic modeling step and a supervised classifier step. The unsupervised modeling step can be also considered as a feature extraction step in the sense that its output can be fed into a supervised classifier subsequently. Specifically, we use the posterior Dirichlet parameter which represents the probability distribution over latent acoustic topics as the feature vector of the corresponding audio clip assuming that audio content under the same genre would have similar latent acoustic topic distributions. For the supervised classifier step, we utilize a Support Vector Machine (SVM) with Bhattacharyya kernel [17] as the machine learning algorithm. Since the SVM was originally designed for binary classification tasks, we use a one-against-one structure for a multi-class classifier.

As a comparison baseline, we use a conventional GMM method; it models the distribution of feature vectors with a given set of Gaussian distributions with respect to their corresponding classes. It has a somewhat similar concept with ATM in the sense that it models the distribution of feature vectors using a number of latent components, i.e., Gaussians distributions. Therefore, for fair comparisons with ATM, we set the number of Gaussian mixtures same with the number of latent acoustic topics (in this work, 64). Note that the feature vectors can be directly used without the quantization process in the GMM.

3. EXPERIMENTS

3.1. Database

We use the RAI dataset [4] which is one of the largest available datasets for TV genre classification tasks and has been used as a benchmark dataset [4, 5, 6]. It contains various TV programs broadcasted in Italian TV channels, namely RAI-1, RAI-2 and RAI-3. The TV programs can be categorized into seven different genres, i.e., *cartoons, commercials, football, music, news, talk shows,* and *weather forecasts*. Each TV program includes a stereo audio signal (2 channels) whose sampling rate is 16 kHz. We mix down the two channels into one channel before further processing. Table 1 shows the statistics of the RAI dataset; see [4] for more details.

For consistency with other benchmarking works, we perform a 6-fold cross-validation; we partition the database into six exclusive subsets randomly in a way that the numbers of clips for individual classes are as equivalent as possible. In each fold, we retain one subset for testing while using the rest of the five subsets for training. All training procedures for the corresponding fold, such as building an acoustic dictionary, modeling acoustic topics, and training SVM classifiers, are done using the training subsets.

Table 1. Number of TV programs and their total length of different genres included in the RAI dataset.

Genre	Number of programs	Total length (in minutes)
Cartoon (CT)	27	433
Commercial (CM)	58	184
Football (FB)	22	1061
Music show (MU)	7	36
News (NE)	49	1039
Talk show (TS)	39	1299
Weather Forecast (WF)	60	112
Total	262	4167

3.2. Experimental Setup

We perform experiments under two different scenarios: on-line and off-line. The off-line system takes advantage of the boundary information of the TV programs, while the on-line the system does not have prior knowledge about boundaries so those need to be detected. In this work, for simplicity, we design the system to yield classification results within fixed length of segments to simulate the on-line scenario.

Note that we use the same models in both on-line and off-line scenarios as shown in Fig. 2. At the training phase, boundaries are always assumed known, but at test phase there are two different test conditions: off-line assuming known boundaries and on-line assuming just streaming program data in which case the boundaries are not known and classification is done on a segment by segment basis. The segment length varies from 0.5 to 6 seconds with an increment of 0.5 seconds.

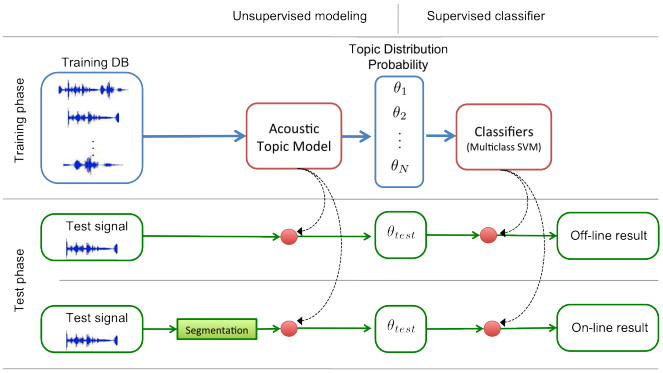


Fig. 2. Diagram of the experimental setup.

To evaluate the performance, we use both *accuracy* and *F-measure*. While accuracy provides a general idea of the performance (weighted mean of per-class *recall* values), F-measure considers both *precision* and *recall* and can be written as a harmonic mean of those two, i.e.,

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

where

$$precision = \frac{\text{number of correctly classified trials in class C}}{\text{total number of test trials classified as class C}}$$

$$recall = \frac{\text{number of correctly classified trials in class C}}{\text{total number of test trials from class C}} \quad .$$

4. RESULTS AND DISCUSSIONS

We start by providing results for the off-line task. Table 2 shows the results of the off-line task in terms of accuracy compared with other benchmarks. Since most of previous work has focused on multimodal approaches which consider both audio and visual content, it is difficult to get results using only audio information, especially in [4] and [5]. However, as shown in the table, deploying only audio content with either GMM or ATM can yield performances competitive with other audio-visual based approaches, especially with [4] and [5], and outperforms the audio-only GMM-based approach of [6].

Fig. 3 depicts the per-class F-measure of the off-line tasks. It is notable that the F-measure of *music show* using ATM is much lower than the one using GMM. It maybe due to unbalanced data distribution of the database; as shown in Table 1, there are only 7 clips for *music show* which indicates that each fold in the cross-validation framework contains one or two instances. This may play an adverse role against ATM rather than GMM since ATM generates only one segment-level feature vector for each segment, while GMM utilizes a considerable amount of frame-level feature vectors.

To provide further investigation, we show confusion matrices in Table 3. Table 3(a) and 3(b) represent the confusion matrices of off-line tasks using using GMM and ATM, respectively. The rows represent the ground-truth classes while the columns represent the predicted classes. The rows are normalized such that the sum in each row equals 1; the diagonal terms, therefore, are per-class recall values. It is clear from the tables that detecting cartoon, commercial and football from others is perfect, as far as this database is concerned. One can also observe that there exist non-trivial confusions between *news*, *talk show* and *weather forecast*. What is notable is the very low recall value of *music show* with ATM (more than half of *music show* instances are misclassified into *commercial*) and relatively low recall value of *talk show* for GMM (it is often misclassified as *news*). These types of mis-classification (*music* as *commercial* or *talk show* as *news*) may be allowable in some applications since they

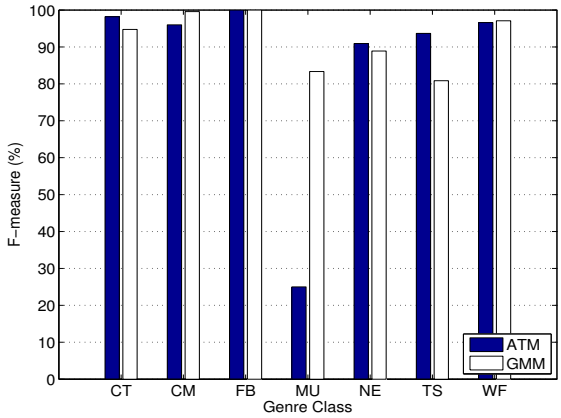


Fig. 3. Per-class F-measure for off-line tasks.

Table 2. Off-line results in terms of accuracy compared with benchmarks.

Accuracy (%)	[4]	[5]	[6]	GMM	ATM
Audio only	-	-	86.6	93.6	94.3
Audio-Visual	92.0	94.9	99.6	-	-

Table 3. Confusion matrix of off-line tasks using (a) GMM and (b) ATM methods. The rows represent the ground-truth classes while the columns represent the predicted classes. The rows are normalized such that the sum in each row is 1, thus the diagonal terms are per-class recall values. Blank elements represent zeros.

(a) GMM

	CT	CM	FB	MU	NE	TS	WF
CT	1.00						
CM		1.00					
FB			1.00				
MU	0.29			0.71			
NE					0.94	0.06	
TS					0.21	0.73	0.05
WF	0.01	0.01			0.01		0.98

(b) ATM

	CT	CM	FB	MU	NE	TS	WF
CT	1.00						
CM		1.00					
FB			1.00				
MU	0.14	0.57		0.14	0.14		
NE					0.92	0.06	0.02
TS					0.05	0.95	0.05
WF		0.02			0.03		0.95

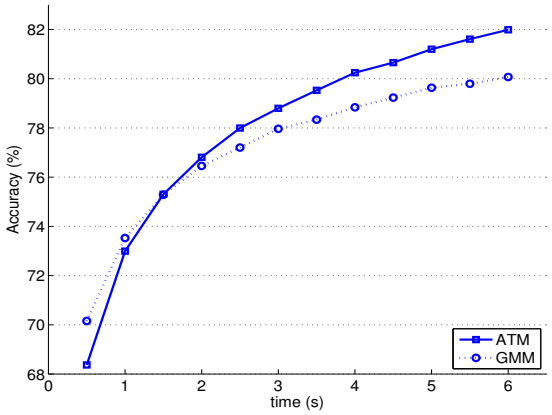


Fig. 4. Classification accuracy for short segments according to the length of segments.

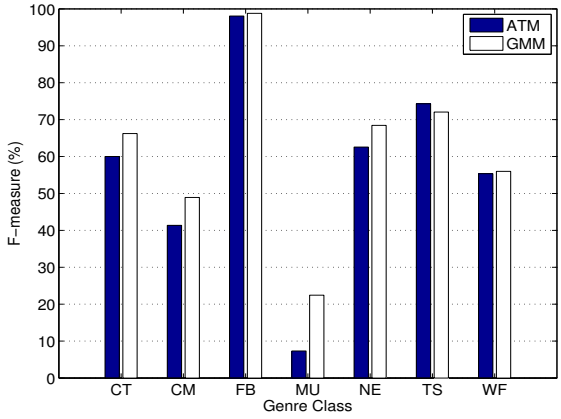
share common properties.

Now we report the performance of on-line genre classification tasks. It is worth reminding that the models are the same as the off-line scenario since we assume that the training data is available even in the on-line scenario. Fig. 4 illustrates the performance of tasks using ATM and GMM in terms of accuracy according to the length of segments. As shown in the figure, the performance increases with rising segment lengths in both the ATM and GMM case. We can observe, however, that GMM outperforms ATM if the length of segments is shorter or equal to 1 second, although performance drops significantly with such frame lengths. This is reasonable since ATM tries to model context, while GMM models content-based distribution of feature vectors, hence it degrades if the data length is so short that it does not represent the modeled context.

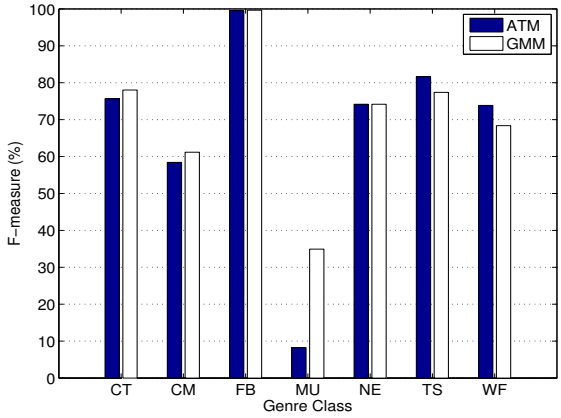
Per-class F-measure of the on-line tasks is also provided in Fig. 5: (a) for 1-second-long segments and (b) for 6-second-long segments. One can easily observe that the performance considering *music show* using ATM is significantly lower than other classes. This is expected based on the observations above, which were made using the same trained models but different test conditions. It is also interesting to observe that the segments from *football* are rarely misclassified even with very short segments. It might be due to its unique background noise conditions, i.e., a noise from a crowd, which is hardly found in other classes hence allowing classification even with minimal context.

5. CONCLUSION AND FUTURE WORK

In this work, we investigated on-line genre classification of TV programs using audio content. Particularly, we deployed an acoustic topic model (ATM) to capture contextual information embedded within audio segments for this purpose. We investigated both off-line (when the entire audio clip is assumed to be available for classification) and on-line scenarios (where data is assumed to be streaming and classification is done on a short term basis) using a 110 hours Italian TV broadcast database (262 programs). The off-line experimental results suggest that the proposed method using just audio content yields competitive performance (94.3% accuracy) with other benchmarks using audio-visual features (92.0% in [4] and



(a) 1 second



(b) 6 seconds

Fig. 5. Per-class F-measure for short segments: (a) 1 second and (b) 6 seconds.

94.9% in [5]) and outperforms conventional audio-based approaches (86.6% in [6]). The results in on-line tasks showed promising results in classifying genres of TV programs with short segments (73% accuracy for 1 second-long segments) and also suggested that ATM performs better than conventional GMM when the length of audio segments is longer.

By investigating the confusion matrices and per-class F-measure, we found that detecting *music show* from others is very challenging. In the future, therefore, we will study on methodologies that can detect *music show* from others successfully, e.g., structure- and harmony- related features. Extra data collection of music shows might be needed for appropriate analysis. We will also study on automatic segmentation methods so that we can automatically detect when to yield the classification results, as done in [18].

6. ACKNOWLEDGEMENT

The authors would like to thank Dr. Messina and Dr. Montagnuolo from RAI Centre for Research and Technological Innovation for providing the TV program data.

7. REFERENCES

[1] M. Barbieri, P. Fonseca, M. A. Peters, and L. Wang, "Multimedia content analysis for consumer electronics," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, ser. CIVR '08. New York, NY, USA: ACM, 2008, pp. 601–608. [Online]. Available: http://bib-ezproxy.epfl.ch:2512/10.1145/1386352.1386439

[2] P. Ferguson, C. Gurrin, H. Lee, S. Sav, A. F. Smeaton, N. E. O'Connor, Y.-H. Choi, and H. Park, "Enhancing the functionality of interactive TV with content-based multimedia analysis," in *Proceedings of the 2009 11th IEEE International Symposium on Multimedia*, ser. ISM '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 495–500. [Online]. Available: http://dx.doi.org/10.1109/ISM.2009.70

[3] Multimedia grand challenge 2009. [Online]. Available: http://comminfo.rutgers.edu/conferences/mmchallenge/

[4] M. Montagnuolo and A. Messina, "TV genre classification using multimodal information and multilayer perceptrons," in *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence on AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*, ser. AI*IA '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 730–741.

[5] ——, "Parallel neural networks for multimodal video genre classification," *Multimedia Tools and Applications*, vol. 41, pp. 125–159, 2009, 10.1007/s11042-008-0222-3. [Online]. Available: http://dx.doi.org/10.1007/s11042-008-0222-3

[6] H. Ekenel and T. Semela, "Multimodal genre classification of TV programs and youtube videos," *Multimedia Tools and Applications*, pp. 1–21, 2011, 10.1007/s11042-011-0923-x. [Online]. Available: http://dx.doi.org/10.1007/s11042-011-0923-x

[7] K. Lee and D. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1406–1416, 2010.

[8] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)*, 1999, pp. 50–57.

[9] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2009, pp. 37–40.

[10] S. Kim, P. Georgiou, and S. Narayanan, "Latent acoustic topic models for unstructured audio classification," *APSIPA Transactions of Signal and Information Processing*, vol. 1, Nov. 2012.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Advances in Neural Information Processing Systems*, 2007, pp. 121–128.

[13] S. Kim, P. G. Georgiou, and S. Narayanan, "Supervised acoustic topic model for unstructured audio information retrieval," in *Asia Pacific Signal and Information Processing Association (APSIPA) annual summit and conference*, 2010.

[14] S. Kim, P. Georgiou, and S. Narayanan, "Supervised acoustic topic model with a consequent classifier for unstructured audio classification," in *Workshop on Content-Based Multimedia Indexing (CBMI)*, Jun. 2012, pp. 121–126.

[15] L. R. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[16] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[17] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, December 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=1005332.1016786

[18] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 441–457, 2001.